ManagerTower: Aggregating the Insights of Uni-Modal Experts for Vision-Language Representation Learning

Xiao Xu[1,3], Bei Li[2,3], Chenfei Wu[3], Shao-Yen Tseng[4], Anahita Bhiwandiwalla[4], Shachar Rosenman[4], Vasudev Lal[4], Wanxiang Che[1], Nan Duan[3]

[1]Harbin Institute of Technology, [2]Northeastern University, [3]Microsoft Research Asia, [4]Intel Labs

ACL 2023 Oral || Presenter: Xiao Xu || July, 2023









Work done during the internship of Microsoft Research Asia NLC Group.

Outline

- Background
- Motivation
- Architecture & Manager Design
- Visualization

Background: Vision-Language Learning

What is Vision-Language (VL) Learning?

Goal: Train a smart AI system that can understand both image and text.

Approach: Large-scale self-supervised pre-training on image-text pairs.

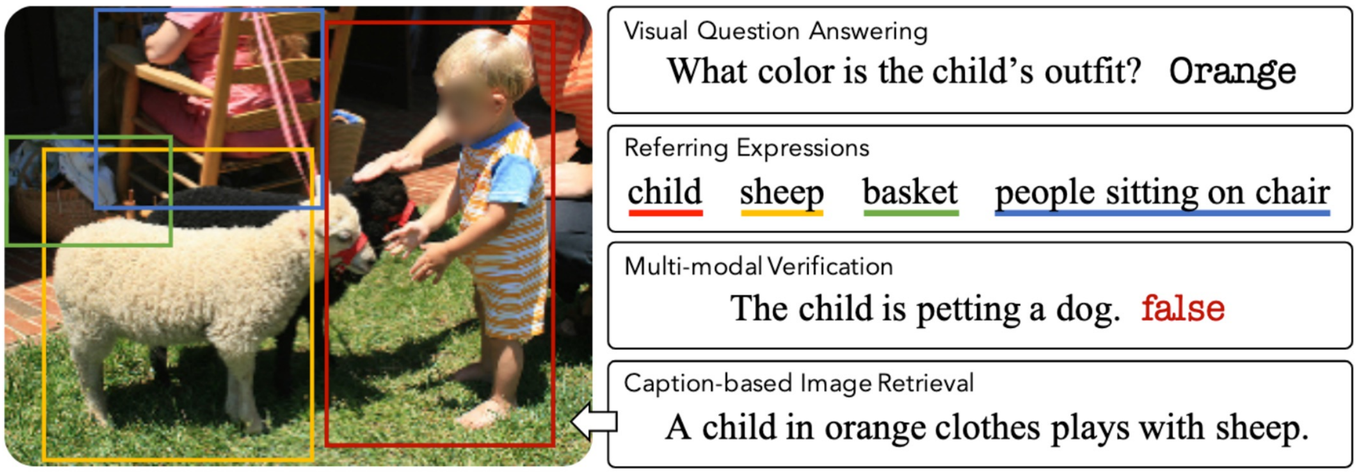
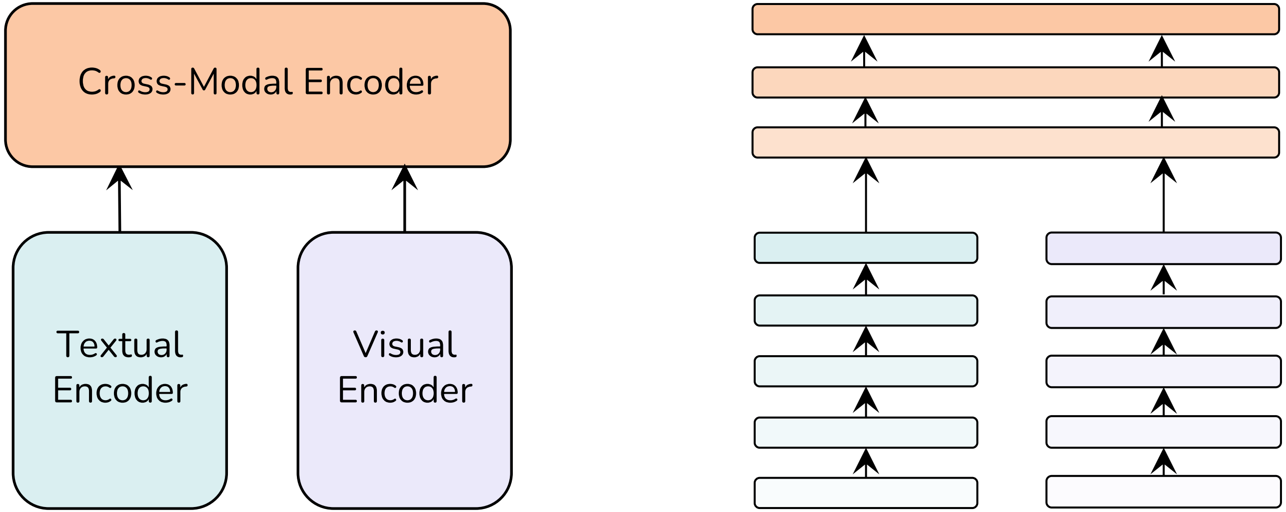


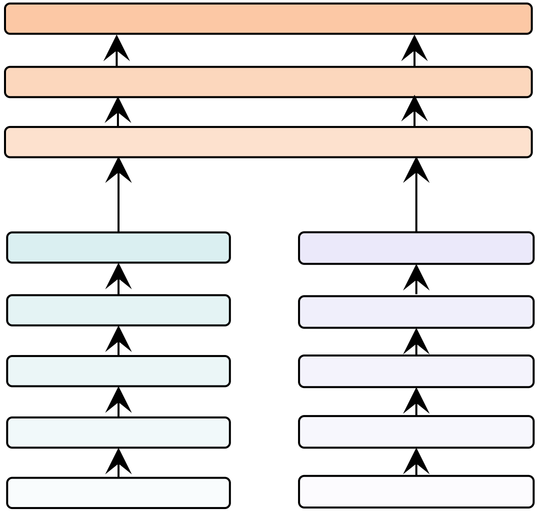
Image from: https://arxiv.org/abs/1912.02315.

Two-Tower Architecture



Two-Tower

Two-Tower vs. BridgeTower



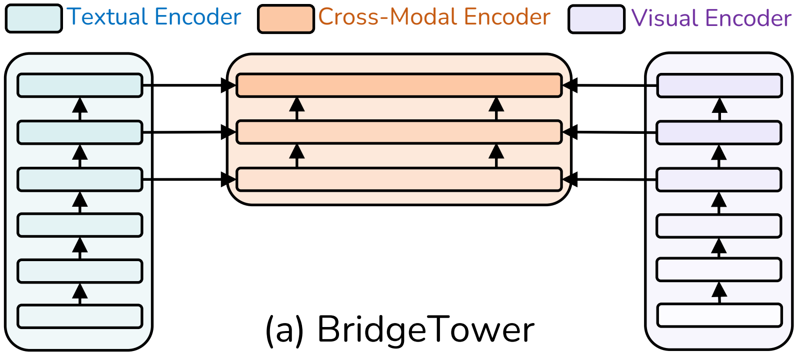
Two-Tower

BridgeTower

Motivation: Adaptively Exploit Uni-Modal Insights

Limitations of BridgeTower

- Ineffective layer-by-layer utilization
- The number of cross-modal layers is tied to the number of uni-modal layer representations it used



(a) BridgeTower

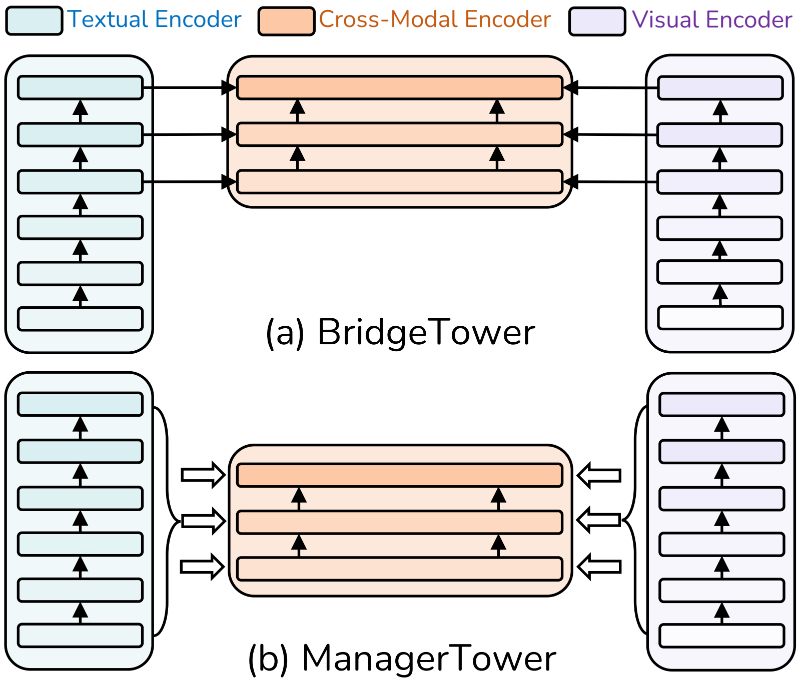
BridgeTower vs. ManagerTower

Limitations of BridgeTower

- Ineffective layer-by-layer utilization
- The number of cross-modal layers is tied to the number of uni-modal layer representations it used

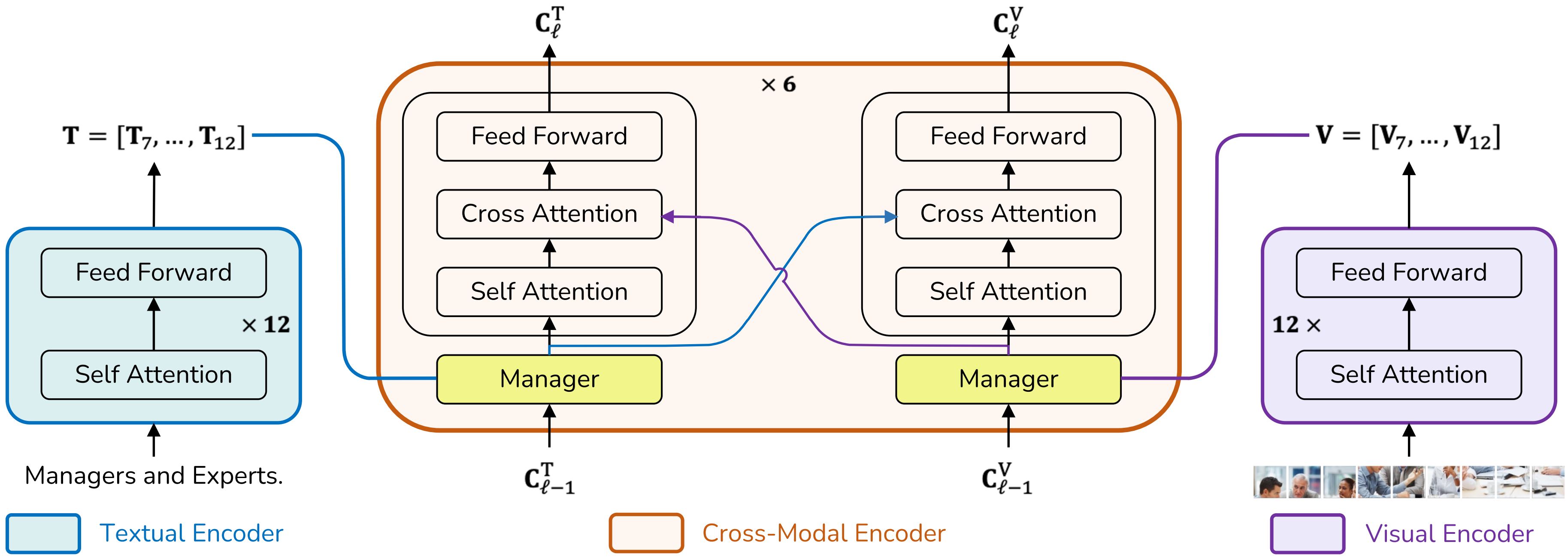
Advances of ManagerTower

- Takes multi-layer uni-modal representations as the insights of pre-trained uni-modal experts at different levels
- Adaptively aggregates insights via managers in each cross-modal layer



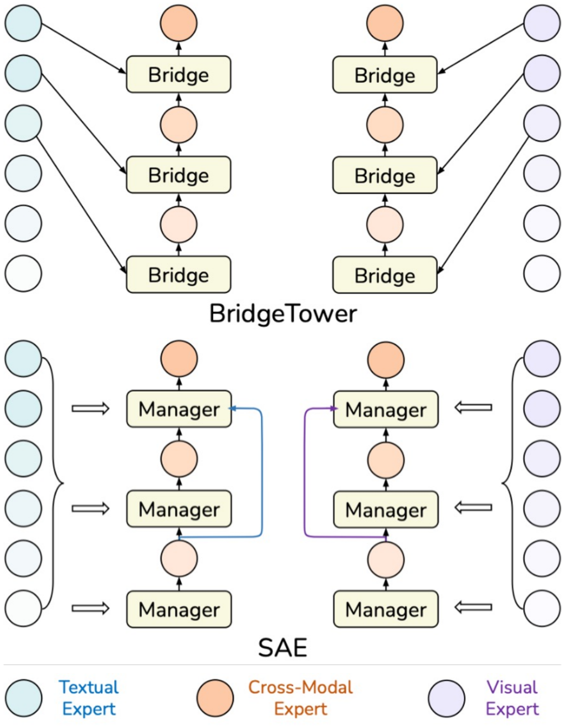
Architecture & Manager Design

ManagerTower Architecture



ManagerTower can work with any visual, textual, or cross-modal encoder.

Static Aggregation of Experts (SAE) Manager

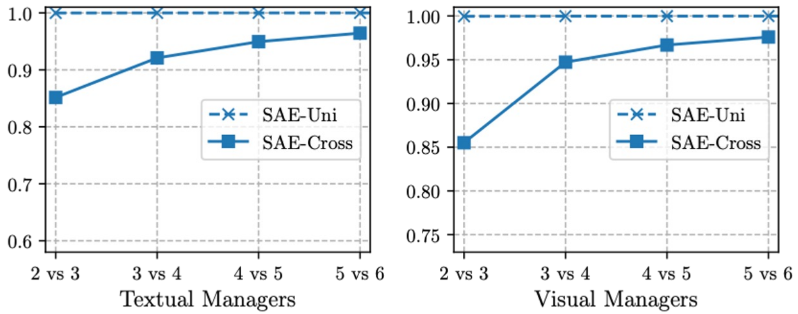


ℓ: cross-modal layer index

$$\text{Input}\begin{cases}\text{Uni-Modal Part:} & \mathrm{L}\times\mathrm{D}\\ \text{Cross-Modal Part:} & \mathrm{L}\times\mathrm{D}\end{cases}\xrightarrow{\text{Bridge}}\text{Output}:\mathrm{L}\times\mathrm{D} \quad (1)$$

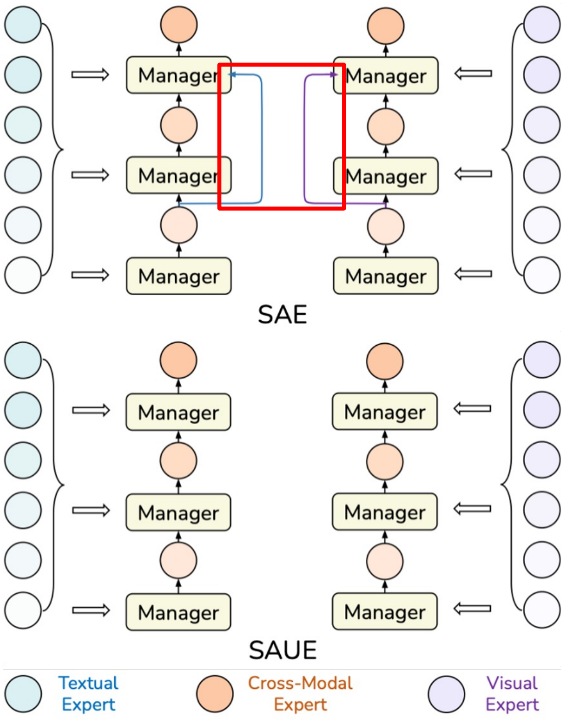
$$\text{Input}\begin{cases}\text{Uni-Modal Part:} & 6\times\mathrm{L}\times\mathrm{D}\\ \text{Cross-Modal Part:} & (\boxed{\ell}-1)\times\mathrm{L}\times\mathrm{D}\\ \text{Learned Weights:} & (6+\ell-1)\times\mathrm{D}\end{cases}\xrightarrow{\text{Manager}}\boxed{\begin{cases}\text{Uni-Modal Aggregated:} & \mathrm{L}\times\mathrm{D}\\ \text{Cross-Modal Aggregated:} & \mathrm{L}\times\mathrm{D}\end{cases}}\xrightarrow{\text{Manager}}\text{Output}:\mathrm{L}\times\mathrm{D} \quad (2)$$

Cosine similarity of aggregated representations between every two consecutive managers



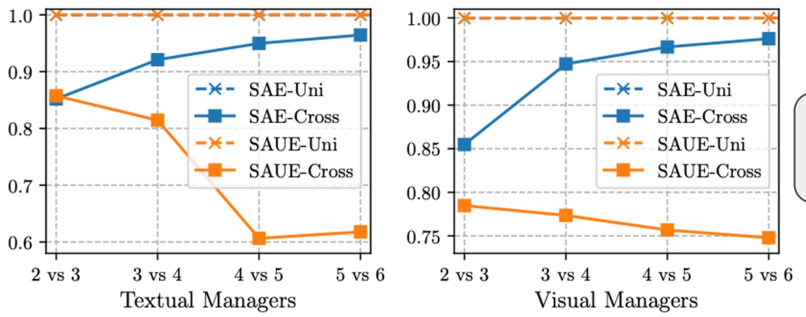
1. Uni-modal similarity ≈ 1
2. Cross-modal similarity increases with depth and gets closer to 1

Static Aggregation of Uni-Modal Experts (SAUE) Manager



$$\text{Input}\begin{cases}\text{Uni-Modal Part:} & 6 \times \mathrm{L} \times \mathrm{D}\\ \text{Cross-Modal Part:} & (\ell - 1) \times \mathrm{L} \times \mathrm{D}\\ \boxed{\text{Learned Weights:} \quad (6 + \ell - 1) \times \mathrm{D}}\end{cases} \xrightarrow{\text{Manager}} \begin{cases}\text{Uni-Modal Aggregated:} & \mathrm{L} \times \mathrm{D}\\ \text{Cross-Modal Aggregated:} & \mathrm{L} \times \mathrm{D}\end{cases} \xrightarrow{\text{Manager}} \text{Output} : \mathrm{L} \times \mathrm{D} \quad (2)$$

$$\text{Input}\begin{cases}\text{Uni-Modal Part:} & 6 \times \mathrm{L} \times \mathrm{D}\\ \text{Cross-Modal Part:} & \mathrm{L} \times \mathrm{D}\\ \boxed{\text{Learned Weights:} \quad 7 \times \mathrm{D}}\end{cases} \xrightarrow{\text{Manager}} \begin{cases}\text{Uni-Modal Aggregated:} & \mathrm{L} \times \mathrm{D}\\ \text{Cross-Modal Aggregated:} & \mathrm{L} \times \mathrm{D}\end{cases} \xrightarrow{\text{Manager}} \text{Output} : \mathrm{L} \times \mathrm{D} \quad (3)$$

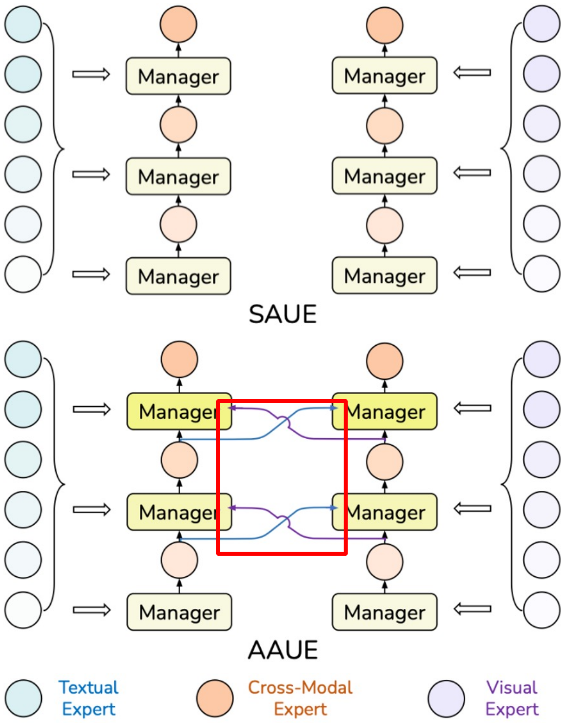


Cross-modal similarity decreases with depth

1. Uni-modal similarity still ≈ 1
2. Input-independent learned weights: N × D

Intuition: the need for uni-modal semantic knowledge varies among cross-modal layers, tokens and samples.

Adaptive Aggregation of Uni-Modal Experts (AAUE) Manager



$$\text{Input}\begin{cases}\text{Uni-Modal Part:} & 6\times \mathrm{L}\times \mathrm{D}\\ \text{Cross-Modal Part:} & \mathrm{L}\times \mathrm{D}\\ \boxed{\text{Learned Weights:} \quad 7\times \mathrm{D}}\end{cases} \xrightarrow{\text{Manager}} \begin{cases}\text{Uni-Modal Aggregated:} & \mathrm{L}\times \mathrm{D}\\ \text{Cross-Modal Aggregated:} & \mathrm{L}\times \mathrm{D}\end{cases} \xrightarrow{\text{Manager}} \text{Output}: \mathrm{L}\times \mathrm{D} \quad (3)$$

$$\text{Input}\begin{cases}\text{Uni-Modal Part:} & 6\times \mathrm{L}\times \mathrm{D}\\ \text{Cross-Modal Part:} & \mathrm{L}\times \mathrm{D}\\ \boxed{\text{Generated Weights:} \quad 7\times \mathrm{L}}\end{cases} \xrightarrow{\text{Manager}} \begin{cases}\text{Uni-Modal Aggregated:} & \mathrm{L}\times \mathrm{D}\\ \text{Cross-Modal Aggregated:} & \mathrm{L}\times \mathrm{D}\end{cases} \xrightarrow{\text{Manager}} \text{Output}: \mathrm{L}\times \mathrm{D} \quad (4)$$

SAE & SAUE managers: learned weights, static sentence-level managers
AAUE managers: generated weights, adaptive token-level managers

Type	Visual Query	Weight	Test-Dev	$\mathrm{R_{MEAN}}$
BridgeTower	-	$\mathrm{N}\times 1$	75.91	93.33
SAE	-	$\mathrm{N}\times 1$	76.19	93.57
	-	$\mathrm{N}\times \mathrm{D}$	76.18	93.73
SAUE	-	$\mathrm{N}\times 1$	76.38	93.75
	-	$\mathrm{N}\times \mathrm{D}$	76.55	93.82
AAUE	$\mathbf{C}^{\mathrm{V}}_{\ell-1}$	$\mathrm{N}\times \mathrm{L}$	76.52	93.84
	$\boxed{\mathrm{CA}(\mathbf{C}^{\mathrm{V}}_{\ell-1}, \mathbf{C}^{\mathrm{T}}_{\ell-1})}$	$\mathrm{N}\times \mathrm{L}$	**76.65**	**93.97**

CA: Cross-Attention

Cross-Modal Fused Query: $\mathrm{L_V}\times \mathrm{D}, \mathrm{L_T}\times \mathrm{D} \xrightarrow{\text{Cross-Attention}} \mathrm{L_V}\times \mathrm{D}$

AAUE managers achieves best performance.

Main Results

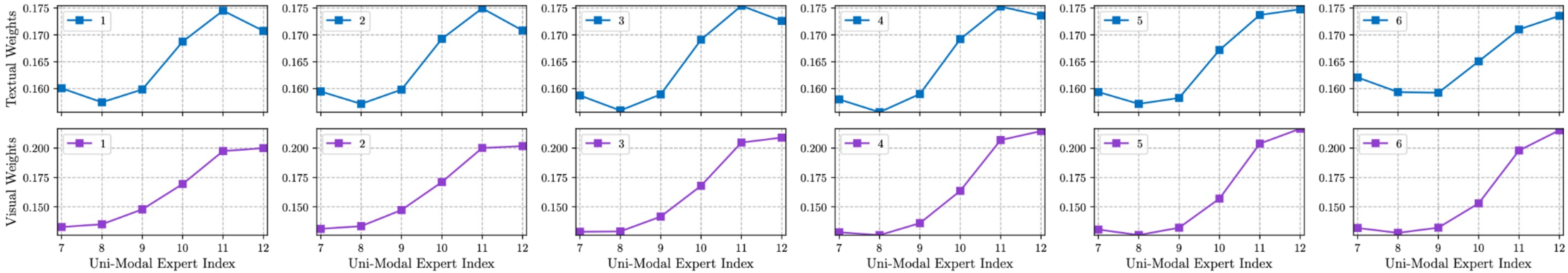
Model	# Pre-train Images	Visual Backbone	VQAv2 Test-Dev	VQAv2 Test-Std	SNLI-VE Dev	SNLI-VE Test	NLVR^2 Dev	NLVR^2 Test-P	Flickr30K IR@1	Flickr30K TR@1
Base-size models pre-trained on 4M public data										
ViLT$_{\text{BASE}}$ (Kim et al., 2021)	4M	ViT-B-384/32	71.26	-	-	-	75.70	76.13	64.4	83.5
UNITER$_{\text{BASE}}$ (Chen et al., 2020) $*$	4M	Faster R-CNN	72.70	72.91	78.59	78.28	77.18	77.85	72.52	85.90
VILLA$_{\text{BASE}}$ (Gan et al., 2020) $*$	4M	Faster R-CNN	73.59	73.67	79.47	79.03	78.39	79.30	74.74	86.60
UNIMO$_{\text{BASE}}$ (Li et al., 2021b)	4M	Faster R-CNN	73.79	74.02	80.00	79.10	-	-	74.66	89.70
ALBEF$_{\text{BASE}}$ (Li et al., 2021a) $*$	4M	DeiT-B-224/16	74.54	74.70	80.14	80.30	80.24	80.50	82.8	94.3
VinVL$_{\text{BASE}}$ (Zhang et al., 2021)	5.7M	ResNeXt-152	75.95	76.12	-	-	82.05	83.08	-	-
METER-Swin$_{\text{BASE}}$ (Dou et al., 2022)	4M	Swin-B-384/32	76.43	76.42	80.61	80.45	82.23	82.47	79.02	92.40
VLMo$_{\text{BASE}}$ (Wang et al., 2021a)	4M	BEiT-B-224/16	76.64	76.89	-	-	82.77	83.34	79.3	92.3
METER-CLIP$_{\text{BASE}}$ (Dou et al., 2022)	4M	CLIP-ViT-B-224/16	77.68	77.64	80.86	81.19	82.33	83.05	82.22	94.30
BridgeTower$_{\text{BASE}}$ (Xu et al., 2022)	4M	CLIP-ViT-B-224/16	78.66	78.73	81.11	81.19	81.85	83.09	85.83	94.73
ManagerTower$_{\text{BASE}}$ **(Ours)**	4M	CLIP-ViT-B-224/16	**79.39**	**79.15**	**81.26**	**81.44**	**82.81**	**83.34**	**86.56**	**95.64**
Models pre-trained on more data and/or with larger size										
UNITER$_{\text{LARGE}}$ (Chen et al., 2020) $*$	4M	Faster R-CNN	73.82	74.02	79.39	79.38	79.12	79.98	75.56	87.30
VILLA$_{\text{LARGE}}$ (Gan et al., 2020) $*$	4M	Faster R-CNN	74.69	74.87	80.18	80.02	79.76	81.47	76.26	87.90
UNIMO$_{\text{LARGE}}$ (Li et al., 2021b)	4M	Faster R-CNN	75.06	75.27	81.11	80.63	-	-	78.04	89.40
ALBEF$_{\text{BASE}}$ (Li et al., 2021a) $*$	14M	DeiT-B-224/16	75.84	76.04	80.80	80.91	82.55	83.14	85.6	95.9
VinVL$_{\text{LARGE}}$ (Zhang et al., 2021)	5.7M	ResNeXt-152	76.52	76.63	-	-	82.67	83.98	-	-
BLIP$_{\text{BASE}}$ (Li et al., 2022a) $*$	14M	DeiT-B-224/16	77.54	77.62	-	-	82.67	82.30	87.2	96.6
SimVLM$_{\text{BASE}}$ (Wang et al., 2021b) $\star$	1.8B	ResNet-101	77.87	78.14	84.20	84.15	81.72	81.77	-	-
BLIP$_{\text{BASE}}$ (Li et al., 2022a) $*$	129M	DeiT-B-224/16	78.24	78.17	-	-	82.48	83.08	87.3	97.3
SimVLM$_{\text{LARGE}}$ (Wang et al., 2021b) $\star$	1.8B	ResNet-152	79.32	79.56	85.68	85.62	84.13	84.84	-	-
VLMo$_{\text{LARGE}}$ (Wang et al., 2021a)	4M	BEiT-L-224/16	79.94	79.98	-	-	85.64	86.86	84.5	95.3
SimVLM$_{\text{HUGE}}$ (Wang et al., 2021b) $\star$	1.8B	Larger ResNet-152	80.03	80.34	86.21	86.32	84.53	85.15	-	-

Follow METER's and BridgeTower's setting + 4M Vision-Language Pre-training + Managers => significant gains and outperforms some models trained with more data and parameters.

Visualization of Aggregation Weights

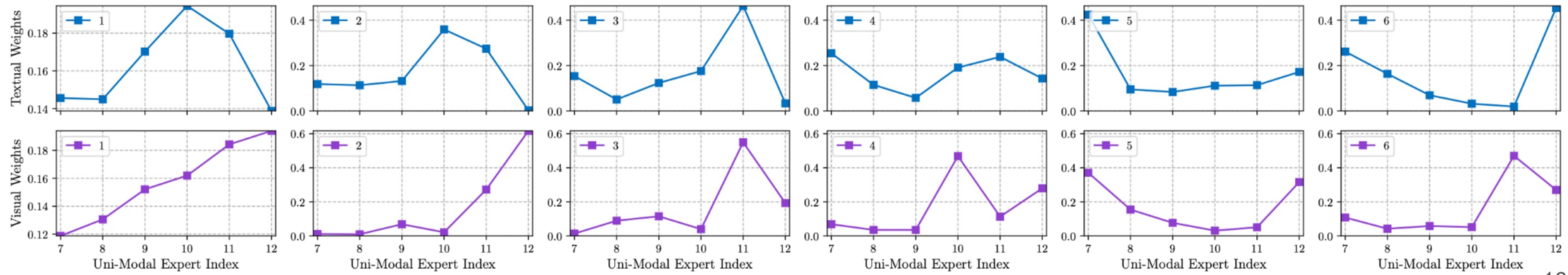
SAUE Managers

Horizontal: similar progressive weight distributions



AAUE Managers

Horizontal: diverse weight distributions



Take-Away Messages

- Introduce managers in each cross-modal layer to
  - adaptively aggregate the insights of pre-trained uni-modal experts at different levels
  - flexibly generate different aggregation weights for different tokens in different samples
  - facilitate more comprehensive cross-modal alignment and fusion
- Cross-modal fused query
  - incorporates the output visual & textual representations of the previous cross-modal layer
  - to help managers to correctly aggregate uni-modal semantic knowledge required by the current cross-modal layer
- ManagerTower can work with any visual, textual, or cross-modal encoder

Thanks & QA

Xiao Xu[1,3], Bei Li[2,3], Chenfei Wu[3], Shao-Yen Tseng[4], Anahita Bhiwandiwalla[4], Shachar Rosenman[4], Vasudev Lal[4], Wanxiang Che[1], Nan Duan[3]

[1]Harbin Institute of Technology, [2]Northeastern University, [3]Microsoft Research Asia, [4]Intel Labs

ACL 2023 Oral || Presenter: Xiao Xu || July, 2023









Slides and more in https://looperxx.github.io/.